

The cross-correlation measure of families of finite binary sequences: limiting distributions and minimal values

László Mériai,

Johann Radon Institute for Computational and Applied Mathematics

Austrian Academy of Sciences

Altenbergerstr. 69, 4040 Linz, Austria

e-mail: merai@cs.elte.hu

August 10, 2016

Abstract

Gyarmati, Mauduit and Sárközy introduced the *cross-correlation measure* $\Phi_k(G)$ of order k to measure the level of pseudorandom properties of families of finite binary sequences. In an earlier paper we estimated the cross-correlation measure of a random family of binary sequences. In this paper, we sharpen these earlier results by showing that for random families, the cross-correlation measure converges strongly, and so has limiting distribution. We also give sharp bounds to the minimum values of the cross-correlation measure, which settles a problem of Gyarmati, Mauduit and Sárközy nearly completely.

2000 Mathematics Subject Classification: Primary 11K45, 68R15, 60C05

Key words and phases: pseudorandom sequences, binary sequence, correlation measure, cross-correlation measure

1 Introduction

Recently, in a series of papers the pseudorandomness of *finite binary sequences* $E_N = (e_1, \dots, e_N) \in \{-1, 1\}^N$ has been studied. In particular, measures of pseudorandomness have been defined and investigated; see [3, 5, 9, 11] and the references therein.

For example, Mauduit and Sárközy [11] introduced the *correlation measure* $C_k(E_N)$ of order k of the binary sequence E_N . Namely, for a k -tuple $D = (d_1, \dots, d_k)$ with non-negative integers $0 \leq d_1 < \dots < d_k < N$ and $M \in \mathbb{N}$ with $M + d_k \leq N$ write

$$V_k(E_N, M, D) = \sum_{n=1}^M e_{n+d_1} \dots e_{n+d_k}.$$

Then $C_k(E_N)$ is defined as

$$C_k(E_N) = \max_{M,D} |V(E_N, M, D)| = \max_{M,D} \left| \sum_{n=1}^M e_{n+d_1} \dots e_{n+d_k} \right|.$$

The final publication is available at Elsevier via <http://dx.doi.org/10.1016/j.dam.2016.06.024>
 © 2016. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

This measure has been widely studied, see, for example [1, 2, 3, 5, 8, 12, 17]. In particular, Alon, Kohayakawa, Mauduit, Moreira and Rödl [3] obtained the typical order of magnitude of $C_k(E_N)$. They proved that, if E_N is chosen uniformly from $\{-1, +1\}^N$, then for all $0 < \varepsilon < 1/16$ the probability that

$$\frac{2}{5}\sqrt{N \log \binom{N}{k}} < C_k(E_N) < \frac{7}{4}\sqrt{N \log \binom{N}{k}}$$

holds for every integer $2 \leq k \leq N/4$ is at least $1 - \varepsilon$ if N is large enough. (Here, and in what follows, we write \log for the natural logarithm, and \log_a for the logarithm to base a .)

They also showed in [3], that the correlation measure $C_k(E_N)$ is concentrated around its mean $\mathbb{E}[C_k]$. Namely, for all $\varepsilon > 0$ and integer function $k = k(N)$ with $2 \leq k \leq \log N - \log \log N$ the probability that

$$1 - \varepsilon < \frac{C_k(E_N)}{\mathbb{E}[C_k]} < 1 + \varepsilon$$

holds is at least $1 - \varepsilon$ if N is large enough.

Recently, K.-U. Schmidt studied the limiting distribution of $C_k(E_N)$ [17]. He showed that if $e_1, e_2, \dots \in \{-1, +1\}$ are chosen independently and uniformly, then for fixed k

$$\frac{C_k(E_N)}{\sqrt{2N \log \binom{N}{k-1}}} \rightarrow 1 \quad \text{almost surely,}$$

as $N \rightarrow \infty$, where $E_N = (e_1, \dots, e_N)$.

Let us now turn to the minimal value of $C_k(E_N)$. Clearly,

$$\min\{C_k(E_N) : E_N \in \{-1, +1\}^N\} = 1 \text{ for odd } k,$$

where the minimum is reached by the alternating sequence $(1, -1, 1, -1, \dots)$. However, for even order, Alon, Kohayakawa, Mauduit, Moreira and Rödl [2] showed that

$$\min\{C_{2k}(E_N) : E_N \in \{-1, +1\}^N\} > \sqrt{\frac{1}{2} \left\lfloor \frac{N}{2k+1} \right\rfloor}, \quad (1)$$

see also [17].

In order to study the pseudorandomness of *families* of finite binary sequences instead of single sequences, Gyarmati, Mauduit and Sárközy [10] introduced the notion of the *cross-correlation measure* (see also the survey paper [15]).

Definition 1. For positive integers N and S , consider a map

$$G_{N,S} : \{1, 2, \dots, S\} \rightarrow \{-1, +1\}^N,$$

and write $G_{N,S}(s) = (e_1(s), \dots, e_N(s)) \in \{-1, 1\}^N$ ($1 \leq s \leq S$).

The *cross-correlation measure* $\Phi_k(G_{N,S})$ of order k of $G_{N,S}$ is defined as

$$\Phi_k(G_{N,S}) = \max \left| \sum_{n=1}^M e_{n+d_1}(s_1) \cdots e_{n+d_k}(s_k) \right|,$$

where the maximum is taken over all integers M, d_1, \dots, d_k and $1 \leq s_1, \dots, s_k \leq S$ such that $0 \leq d_1 \leq d_2 \leq \dots \leq d_k < M + d_k \leq N$ and $d_i \neq d_j$ if $s_i = s_j$.

We remark that in [10] only injective maps $G_{N,S}$ were considered, and the cross-correlation measure is defined for the families $\mathcal{F} = \{G_{N,S}(s) : s = 1, 2, \dots, S\}$ of size S .

The typical order of magnitude of $\Phi_k(G_{N,S})$ was established in [14] for large range of k and for random maps $G_{N,S}$, i.e. when all $e_n(s) \in \{-1, +1\}$ ($1 \leq n \leq N$, $1 \leq s \leq S$) are chosen independently and uniformly.

Theorem 1. *For a given $\varepsilon > 0$, there exists N_0 , such that if $N > N_0$ and $1 \leq \log_2 S < N/12$, then we have with probability at least $1 - \varepsilon$, that*

$$\frac{2}{5} \sqrt{N \left(\log \binom{N}{k} + k \log S \right)} < \Phi_k(G_{N,S}) < \frac{5}{2} \sqrt{N \left(\log \binom{N}{k} + k \log S \right)}$$

for every integer k with $2 \leq k \leq N/(6 \log_2 S)$.

Our first result tells that analogously to the correlation measure of binary sequences, the cross-correlation measure of families $\Phi_k(G_{N,S})$ is concentrated around its mean $\mathbb{E}[\Phi_k(G_{N,S})]$ if k is small enough.

Theorem 2. *For any fixed constant $\varepsilon > 0$ and any integer function $k = k(N)$ with $2 \leq k \leq (\log N + \log S)/\log \log N$, there is a constant $N_0 \geq 12 \log_2 S$ for which the following holds. If $N \geq N_0$, then the probability that*

$$1 - \varepsilon < \frac{\Phi_k(G_{N,S})}{\mathbb{E}[\Phi_k(G_{N,S})]} < 1 + \varepsilon$$

holds is at least $1 - \varepsilon$.

Next, we improve the upper bound in Theorem 1.

Theorem 3. *For positive integers N, S and for $1 \leq s \leq S$ write $G_{N,S}(s) = (e_1(s), e_2(s), \dots, e_N(s))$ ($1 \leq s \leq S$). Let $e_n(s)$ ($1 \leq n \leq N$, $1 \leq s \leq S$) be drawn independently and uniformly at random from $\{-1, 1\}$. For all $\varepsilon > 0$ we have*

$$\mathbb{P} \left[\Phi_k(G_{N,S}) \leq (1 + \varepsilon) \sqrt{2N \log \left(\binom{SN}{k} - \binom{S(N-1)}{k} \right)} \right. \\ \left. \text{for all } k \text{ satisfying } 2 \leq k < SN \right] \rightarrow 1,$$

as $N \rightarrow \infty$.

In order to obtain the asymptotic distribution of the cross-correlation measure $\Phi_k(G_{N,S})$, consider the set Ω of all maps $G_S : \mathbb{N} \rightarrow \{-1, +1\}^{\mathbb{N} \times S}$ and write $G_S(s) = (e_1(s), e_2(s), \dots)$ for $1 \leq s \leq S$. Let us endow Ω the probability measure

$$\mathbb{P}[G_S \in \Omega : e_1(i) = c_{1,i}, e_2(i) = c_{2,i}, \dots, e_N(i) = c_{N,i}, i = 1, \dots, S] = 2^{-NS} \quad (2)$$

for all $N \in \mathbb{N}$ and all $(c_{1,i}, c_{2,i}, \dots, c_{N,i})_{i=1}^S \in \{-1, 1\}^{N \times S}$.

Theorem 4. Let S be a positive integer and let G_S be drawn from Ω equipped with the probability measure defined by (2) with $G_S(s) = (e_1(s), e_2(s), \dots)$ for $1 \leq s \leq S$. Let $G_{N,S}(s) = (e_1(s), \dots, e_N(s))$ for $1 \leq s \leq S$. For a fixed $k \geq 2$ we have

$$\frac{\Phi_k(G_{N,S})}{\sqrt{2N \log \binom{N}{k-1}}} \rightarrow 1 \quad \text{almost surely}$$

as $N \rightarrow \infty$.

Finally, we study the minimum values of the cross-correlation measure $\Phi_k(G_{N,S})$. If $G_{N,S}$ is non-injective, say $G_{N,S}(S-1) = G_{N,S}(S)$, then

$$\Phi_k(G_{N,S}) = \max \{ \Phi_{k-2}(G_{N,S-1}), \Phi_k(G_{N,S-1}) \},$$

thus it is enough to control the minimum values of $\Phi_k(G_{N,S})$ when $G_{N,S}$ is injective.

In [10], Gyarmati, Mauduit and Sárközy showed, that if the order of the measure is odd and S is small, then $\Phi_{2k+1}(G_{N,S})$ can be small.

Proposition 1. Let $N \in \mathbb{N}$, $k, S \in \mathbb{N}$, such that $2k+1 < N$, $S < N$. Then there is an injective map $G_{N,S}$ such that

$$\Phi_{2k+1}(G_{N,S}) \leq 2S.$$

Based on this observation they posed the following problem.

Problem 1. Estimate $\min \{ \Phi_{2k+1}(G_{N,S}) \}$ for any fixed N, k and S , where the minimum is taken over all injective maps $G_{N,S} : \{1, 2, \dots, S\} \rightarrow \{-1, 1\}^N$.

We shall prove

Theorem 5. If k and N are positive integers, then

$$\lfloor \log_2 S - \log_2(2k+1) \rfloor \leq \min \{ \Phi_{2k+1}(G_{N,S}) \} \leq \lceil \log_2 S \rceil,$$

where the minimum is taken over all injective maps $G_{N,S} : \{1, 2, \dots, S\} \rightarrow \{-1, 1\}^N$.

Similarly to the correlation measure, the cross-correlation measure cannot be small if its order is even. From (1) and a trivial estimate we get

$$\Phi_{2k}(G_{N,S}) \geq \max \{ C_{2k}(G_{N,S}(s)) : 1 \leq s \leq S \} \geq \sqrt{\frac{1}{2} \left\lfloor \frac{N}{2k+1} \right\rfloor}.$$

This lower bound can be improved, for example, by essentially a $\log \lfloor S/k \rfloor$ term if S is large.

Theorem 6. If k and N are positive integers, then for all injective maps $G_{N,S} : \{1, 2, \dots, S\} \rightarrow \{-1, 1\}^N$ we have

$$\Phi_{2k}(G_{N,S}) \geq \sqrt{\frac{1}{50} N \log \lfloor S/k \rfloor} \Big/ \log \frac{50N}{\log \lfloor S/k \rfloor} \quad (3)$$

if $2kN \leq S$, and

$$\Phi_{2k}(G_{N,S}) \geq \sqrt{\frac{N}{2 \lfloor k/S \rfloor + 1}} \quad (4)$$

if $2kN > S$.

2 Estimates for $\Phi_k(G_{N,S})$ for random $G_{N,S}$

In this section we shall prove Theorems 2 and 3. The proof of Theorem 2 is based on the following result (see e.g. [13, Lemma 1.2]).

Lemma 2. *Let X_1, \dots, X_n be independent random variables, with X_j taking values in a set A_j for each j . Suppose that the (measurable) function $f : \prod_{j=1}^n A_j \rightarrow \mathbb{R}$ satisfies*

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_j$$

whenever the vectors \mathbf{x} and \mathbf{x}' differ only in the j th co-ordinate. Let Y be the random variable $f(X_1, \dots, X_n)$. Then for any $\theta > 0$,

$$\mathbb{P}[|Y - \mathbb{E}(Y)| \geq \theta] \leq 2 \exp \left\{ -\frac{2\theta^2}{\sum_{j=1}^n c_j^2} \right\}.$$

Lemma 3. *For $\theta \geq 0$ we have*

$$\mathbb{P}[|\Phi_k(G_{N,S}) - \mathbb{E}[\Phi_k(G_{N,S})]| \geq \theta] \leq 2 \exp \left\{ -\frac{\theta^2}{2k^2N} \right\}. \quad (5)$$

Proof. For a fixed $1 \leq j \leq N$ consider two maps $G_{N,S}, G'_{N,S} : \{1, \dots, S\} \rightarrow \{-1, 1\}^N$ with $G_{N,S}(s) = (e_1(s), \dots, e_N(s))$ and $G'_{N,S}(s) = (e'_1(s), \dots, e'_N(s))$ for $1 \leq s \leq S$, such that for all s the sequences $(e_1(s), \dots, e_N(s))$ and $(e'_1(s), \dots, e'_N(s))$ can only differ at the j th position:

$$e_n(s) = e'_n(s) \text{ for } 1 \leq s \leq S \text{ and } n \neq j.$$

Then

$$\left| \sum_{n=1}^M e_{n+d_1}(s_1) \dots e_{n+d_k}(s_k) - \sum_{n=1}^M e'_{n+d_1}(s_1) \dots e'_{n+d_k}(s_k) \right| \leq 2k,$$

therefore

$$|\Phi_k(G_{N,S}) - \Phi_k(G'_{N,S})| \leq 2k$$

which proves (5) by Lemma 2. \square

Theorem 2. By Lemma 3 it is enough to show that if $2 \leq k \leq (\log N + \log S)/\log \log N$, then taking $\theta = \varepsilon \mathbb{E}[\Phi_k(G_{N,S})]$ the right hand side of (5) is $o(1)$. If N is large enough, then by Theorem 1 we have

$$\mathbb{E}[\Phi_k(G_{N,S})] > \frac{1}{5} \sqrt{kN (\log N + \log S)}.$$

Then

$$\frac{\theta^2}{2k^2N} = \frac{\varepsilon^2 (\mathbb{E}[\Phi_k(G_{N,S})])^2}{2k^2N} \geq \frac{\varepsilon^2 \log N + \log S}{50k} \rightarrow \infty,$$

as $N \rightarrow \infty$. \square

Let X_1, \dots, X_N be independent random variables, each taking the values -1 or 1, each with probability 1/2. Define the random variable

$$R_N = \max_{1 \leq m_1 \leq m_2 \leq N} \left| \sum_{j=m_1}^{m_2} X_j \right|.$$

The following lemma states an estimate for large deviation of R_N [17, Lemma 2.2].

Lemma 4. For all $\delta > 0$, there exists $N_0 = N_0(\delta)$ such that for all $N \geq N_0$ and all $\lambda > 2\sqrt{N}$ we have

$$\mathbb{P}[R_N > (1 + \delta)\lambda] \leq \log N \exp\left(-\frac{\lambda^2}{2N}\right).$$

One can obtain in the same way as [3, Claim 18] that the summands in the definition of the cross-correlation measure are pairwise independent.

Lemma 5. Let $1 \leq n, n' \leq N$ be integers, $(s_1, \dots, s_k), (s'_1, \dots, s'_k), (d_1, \dots, d_k)$ and (d'_1, \dots, d'_k) be k -tuples such that $1 \leq s_1, \dots, s_k \leq S$, $1 \leq s'_1, \dots, s'_k \leq S$, $0 \leq d_1 \leq \dots \leq d_k$, $0 \leq d'_1 \leq \dots \leq d'_k$ with $d_i \neq d_j$ if $s_i = s_j$ and $d'_i \neq d'_j$ if $s'_i = s'_j$. If

$$(n, s_1, \dots, s_k, d_1, \dots, d_k) \neq (n', s'_1, \dots, s'_k, d'_1, \dots, d'_k),$$

then

$$e_{n+d_1}(s_1) \cdots e_{n+d_k}(s_k) \quad \text{and} \quad e_{n'+d'_1}(s'_1) \cdots e_{n'+d'_k}(s'_k)$$

are independent.

Throughout the proofs of Theorems 3 and 4 we will frequently use the following well-known bounds to the binomial coefficients

$$\left(\frac{n}{m}\right)^m \leq \binom{n}{m} \leq \left(\frac{en}{m}\right)^m, \quad \text{for } n, m \in \mathbb{N}, \quad 0 < m \leq n. \quad (6)$$

Theorem 3. Write $G_{N,S}(s) = (e_1(s), \dots, e_N(s))$ for $1 \leq s \leq S$. Then writing $d_1 = 0$ we have

$$\Phi_k(G_{N,S}) = \max_{0 \leq d_2 \leq \dots \leq d_k} \max_{s_1, \dots, s_k}^* \max_{1 \leq m_1 \leq m_2 \leq N-d_k} \left| \sum_{n=m_1}^{m_2} e_{n+d_1}(s_1) \cdots e_{n+d_k}(s_k) \right|, \quad (7)$$

where the asterisk indicates that the second maximum is taken over all $1 \leq s_1, \dots, s_k \leq S$ such that $s_i \neq s_j$ if $d_i = d_j$.

Let

$$\lambda = \sqrt{2N \log \left(\binom{SN}{k} - \binom{S(N-1)}{k} \right)}$$

and write $1 + \varepsilon = \sqrt{1 + \gamma}(1 + \delta)$ for some $\gamma, \delta > 0$. By Lemmas 4 and 5 we have

$$\mathbb{P} \left[\max_{1 \leq m_1 \leq m_2 \leq N-d_k} \left| \sum_{n=m_1}^{m_2} e_{n+d_1}(s_1) \cdots e_{n+d_k}(s_k) \right| > (1 + \varepsilon)\lambda \right]$$

is at most

$$\log N \exp \left\{ -\frac{\lambda^2(1 + \gamma)}{2N} \right\} = \frac{\log N}{\left(\binom{SN}{k} - \binom{S(N-1)}{k} \right)^{(1+\gamma)}}$$

if N is large enough.

Summing over all tuples (d_2, \dots, d_k) and (s_1, \dots, s_k) considered in (7) we get

$$\mathbb{P}[\Phi_k(G_N) > (1 + \varepsilon)\lambda] \leq \sum_{0 \leq d_2 \leq \dots \leq d_k} \sum_{s_1, \dots, s_k}^* \frac{\log N}{\left(\binom{SN}{k} - \binom{S(N-1)}{k} \right)^{(1+\gamma)}} \quad (8)$$

if N is large enough.

Denoting the number of zero d_i 's by ℓ , we get that the number of possible tuples is

$$\sum_{\ell=1}^S \binom{S}{\ell} \binom{S(N-1)}{k-\ell} = \binom{SN}{k} - \binom{S(N-1)}{k}, \quad (9)$$

where the equation follows from the Chu–Vandermonde identity

$$\sum_{j=0}^l \binom{m}{j} \binom{n-m}{l-j} = \binom{n}{l}$$

which can be obtained from the coefficient of x^l in the polynomial equation $(1+x)^m(1+x)^{m-n} = (1+x)^n$.

From (8) and (9) we get

$$\begin{aligned} \mathbb{P}[\Phi_k(G_N) > (1+\varepsilon)\lambda] &\leq \frac{\log N}{\left(\binom{SN}{k} - \binom{S(N-1)}{k}\right)^\gamma} \leq \frac{\log N}{\left(\binom{SN}{k} - \binom{S(N-1)}{k}\right)^\gamma} \\ &= \frac{\log N}{\binom{SN-1}{k-1}^\gamma}. \end{aligned}$$

In order to prove the theorem it is enough to show that

$$\sum_{k=2}^{SN-1} \mathbb{P}[\Phi_k(G_N) > (1+\varepsilon)\lambda] \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Let M be an integer such that $M\gamma > 1$. Then, for $N > M/S$ we have that

$$\begin{aligned} \sum_{k=2}^{SN-1} \mathbb{P}[\Phi_k(G_N) > (1+\varepsilon)\lambda] &\leq 2 \sum_{k=1}^{M-1} \frac{\log N}{\binom{SN-1}{k}^\gamma} + 2 \sum_{k=M}^{\lfloor (SN-1)/2 \rfloor} \frac{\log N}{\binom{SN-1}{k}^\gamma} \\ &\leq \frac{2M \log N}{(SN-1)^\gamma} + \frac{SN \log N}{\binom{SN-1}{M}^\gamma} \\ &\leq \frac{2M \log N}{(SN-1)^\gamma} + \frac{M^{M\gamma} \log N}{(SN-1)^{M\gamma-1}}, \end{aligned}$$

using (6). Since $\gamma > 0$ and $M\gamma > 1$, the right hand side tends to zero as $N \rightarrow \infty$ which proves the theorem. \square

3 Limiting distribution

The main ingredient of the proof of Theorem 4 is the following asymptotic result for the mean $\mathbb{E}[\Phi_k(G_{N,S})]$.

Lemma 6. *Let $G_{N,S}(s)$ be drawn independently and uniformly at random from $\{-1, 1\}^N$ for all $1 \leq s \leq S$. Then, as $N \rightarrow \infty$,*

$$\frac{\mathbb{E}[\Phi_k(G_{N,S})]}{\sqrt{2N(k-1)\log N}} \rightarrow 1.$$

Let ℓ and k_1, \dots, k_ℓ be positive integers with $k_1 + \dots + k_\ell = k$. Let $D = (d_1^i, \dots, d_{k_i}^i)_{i=1}^\ell$ be a k -tuple such that

$$0 \leq d_1^i < \dots < d_{k_i}^i \leq \frac{N}{\log N} \quad \text{for } i = 1, \dots, \ell, \quad \text{and} \quad \min_{i=1, \dots, \ell} d_1^i = 0. \quad (10)$$

For distinct $1 \leq s_1, \dots, s_\ell \leq S$ write

$$\begin{aligned} & V_{k_1, \dots, k_\ell}(G_N, s_1, \dots, s_\ell, D) \\ &= \sum_{n=1}^{N - \lfloor \frac{N}{\log N} \rfloor} e_{n+d_1^1}(s_1) \dots e_{n+d_{k_1}^1}(s_1) \dots e_{n+d_1^\ell}(s_\ell) \dots e_{n+d_{k_\ell}^\ell}(s_\ell). \end{aligned}$$

For functions $f(x), g(x)$, we use the standard notation $f(x) \sim g(x)$ to mean $f(x) = g(x)(1 + o(1))$ as $x \rightarrow \infty$.

Lemma 7. *Let $G_{N,S}(s)$ be drawn independently and uniformly at random from $\{-1, 1\}^N$ for all $1 \leq s \leq S$. Then*

$$\begin{aligned} \mathbb{P} \left[|V_{k_1, \dots, k_\ell}(G_{N,S}, s_1, \dots, s_\ell, D)| \geq \sqrt{2N(k-1) \log N} \right] \\ \sim \frac{1}{e^{k-1} N^{k-1} \sqrt{\pi(k-1) \log N}} \end{aligned}$$

as $N \rightarrow \infty$.

We need the following form of the de Moivre-Laplace theorem (see, e.g., [6, Chapter I, Theorem 6]).

Lemma 8. *Let X_1, \dots, X_n be independent random variables, each taking the values -1 or 1 , both with probability $1/2$. For any $c_n > 0$ with $c_n = o(n^{1/6})$ and $c_n \rightarrow \infty$, we have*

$$\mathbb{P} \left[\left| \sum_{i=1}^n X_i \right| \geq c_n \sqrt{n} \right] \sim \sqrt{\frac{2}{\pi}} \frac{1}{c_n} \exp \left\{ -\frac{c_n^2}{2} \right\}.$$

Lemma 7. Write

$$c_N = \sqrt{\frac{2N}{N - \lfloor N/\log N \rfloor} (k-1) \log N}.$$

Then, by Lemmas 5 and 8 we have

$$\begin{aligned} & \mathbb{P} \left[\left| \sum_{n=1}^{N - \lfloor N/\log N \rfloor} e_{n+d_1}(s_1) \dots e_{n+d_k}(s_k) \right| \geq c_N \sqrt{N - \lfloor N/\log N \rfloor} \right] \\ & \sim \frac{1}{\sqrt{\pi(k-1) \log N}} \exp \left\{ -\frac{N}{N - \frac{N}{\log N} + O(1)} (k-1) \log N \right\} \\ & = \frac{1}{e^{k-1} N^{k-1} \sqrt{\pi(k-1) \log N}} e^{-\frac{k-1}{\log N - 1} - O(k \frac{\log N}{N})} \end{aligned}$$

if N is large enough. □

Lemma 9. Let G_S be drawn from Ω with the probability measure defined by (2) and define $G_{N,S}$ as in Theorem 4.

Let ℓ and $k_1, \dots, k_\ell, k'_1, \dots, k'_\ell$ be positive integers with $k_1 + \dots + k_\ell = k'_1 + \dots + k'_\ell = k$, let $1 \leq s_1 < \dots < s_\ell \leq S$ and $D \neq D'$ k -tuples having the form (10). Then writing

$$\lambda = \sqrt{2N(k-1)\log N}$$

we have

$$\begin{aligned} \mathbb{P}[|V_{k_1, \dots, k_\ell}(G_N, s_1, \dots, s_\ell, D)| \geq \lambda \cap |V_{k'_1, \dots, k'_\ell}(G_N, s_1, \dots, s_\ell, D')| \geq \lambda] \\ \leq \frac{23}{N^{2(k-1)}}. \end{aligned}$$

In order to prove Lemma 9 we use the following notation. A tuple (x_1, \dots, x_{2m}) is t -even if there exists a permutation σ of $\{1, 2, \dots, 2m\}$ such that $x_{\sigma(2i-1)} = x_{\sigma(2i)}$ for each $i \in \{1, \dots, t\}$ and t is the largest integer with this property. An m -even tuple is just called *even*.

The following lemma gives an upper bound to the number of even tuples [16, Lemma 2.4].

Lemma 10. Let m and q be positive integers. Then the number of even tuples in $\{1, \dots, m\}^{2q}$ is at most $(2q-1)!!m^q$, where the $(2q-1)!!$ semi-factorial is defined as

$$(2q-1)!! = \frac{(2q)!}{q!2^q} = (2q-1) \cdot (2q-3) \cdots 3 \cdot 1.$$

The following result is an extension of [17, Lemma 3.7].

Lemma 11. Let N , q and t be positive integers satisfying $0 \leq t < q$. Let D and D' be two k -tuples satisfying $D \neq D'$ and (10).

If (x_1, \dots, x_{2q}) is d -even for some $d < q - t$, then the number of $4q$ -tuples $(x_1, \dots, x_{2q}, y_1, \dots, y_{2q})$ in $\{1, \dots, N\}^{4q}$ such that for each $1 \leq s \leq S$ the tuple

$$(x_i + d_1^s, \dots, x_i + d_{k_s}^s, y_i + d_1'^s, \dots, y_i + d_{k'_s}^s)_{i=1}^{2q}$$

is even, is at most

$$(4kq-1)!!N^{2q-(t+1)/3}.$$

Proof. Since the proof is similar to the proof of [17, Lemma 3.7], we leave some details to the reader.

We construct a set of tuples that contains the required $4q$ -tuples as a subset. For each $1 \leq s \leq S$, arrange the $4(k_s + k'_s)q$ variables

$$x_i + d_j^s, y_i + d_l'^s \quad \text{for } 1 \leq i \leq 2q, 1 \leq j \leq k_s, 1 \leq l \leq k'_s \quad (11)$$

into $2(k_s + k'_s)q$ unordered pairs $\{a_1^s, b_1^s\}, \dots, \{a_{2(k_s+k'_s)q}^s, b_{2(k_s+k'_s)q}^s\}$ such that there are at most $k(q-t-1)$ pairs of form $\{x_i + d_j^s, x_{i'} + d_j^s\}$. This can be done in at most $(4kq-1)!!$ ways. We formally set $a_i^s = b_i^s$ for all $1 \leq i \leq 2q$ and all $1 \leq s \leq S$. If this assignment does not yield a contradiction, then we call the arrangement (11) *consistent*.

If there is a pair of form $\{x_i + d_j^s, x_{i'} + d_l^s\}$ with $j \neq l$ in a consistent arrangement, then $i \neq i'$ and x_i determines $x_{i'}$. Likewise, if there is a pair of form $\{y_i + d_j'^s, y_{i'} + d_l'^s\}$ with $j \neq l$ in a consistent arrangement, then $i \neq i'$ and y_i determines $y_{i'}$. On the other hand, if a consistent

arrangement consists a pair of form $\{x_i + d_j^s, y_{i'} + d_l'^s\}$, then x_i determines $y_{i'}$ and at least one other variable in the list

$$x_1, \dots, x_{2q}, y_1, \dots, y_{2q}. \quad (12)$$

Indeed, a consistent arrangement cannot contain pairs involving only the variables

$$x_i + d_1^s, \dots, x_i + d_{k_s}^s, y_{i'} + d_1'^s, \dots, y_{i'} + d_{k'_s}'^s \quad \text{for each } 1 \leq s \leq S.$$

Since for all $1 \leq s \leq S$, $0 \leq d_1^s < \dots < d_{k_s}^s$, and $0 \leq d_1'^s < \dots < d_{k'_s}'^s$, the only possibility for such pairs would be

$$\{x_i + d_1^s, y_{i'} + d_1'^s\}, \dots, \{x_i + d_{k_s}^s, y_{i'} + d_{k'_s}'^s\}, \quad 1 \leq s \leq S. \quad (13)$$

Let u, v be two indices such that $d_1^u = 0$ and $d_1'^v = 0$. Then $x_i = y_{i'} + d_1'^u$ and $x_i + d_1^v = y_{i'}$ so we have $d_1^u = d_1'^v = 0$, thus $x_i = y_{i'}$ and $d_j^u = d_j'^u$ ($j = 1, \dots, k_u$), $d_j^v = d_j'^v$ ($j = 1, \dots, k_v$). Moreover it also follows from (13) that $d_j^s = d_j'^s$ ($j = 1, \dots, k_s$) for $s \neq u, v$, so $D = D'$, a contradiction.

Now, by assumption, each consistent arrangement contains at most $k(q-t-1)$ pairs of form $\{x_i + d_j^s, x_{i'} + d_j'^s\}$ and at most kq pairs of the form $\{y_i + d_j'^s, y_{i'} + d_j'^s\}$, and so at most

$$q - t - 1 + q + \frac{1}{3}(2t + 2) = 2q - \frac{1}{3}(t + 1)$$

of the variables in (12) can be chosen independently. We assign to each of these a value of $\{1, \dots, N\}$. In this way, we construct a set of at most $(4kq-1)!!N^{2q-(t+1)/3}$ tuples that contains the required $4q$ -tuples. \square

Lemma 12. *Let S and k be integers. Let $G_{N,S}(s)$ be drawn independently and uniformly at random from $\{-1, 1\}^N$ for all $1 \leq s \leq S$. Let ℓ and $k_1, \dots, k_\ell, k'_1, \dots, k'_\ell$ be positive integers with $k_1 + \dots + k_\ell = k'_1 + \dots + k'_\ell = k$, let $s_1, \dots, s_\ell \in \mathcal{S}$ distinct elements and let $D \neq D'$ k -tuples having the form (10).*

For $0 \leq h < p$ we have

$$\begin{aligned} \mathbb{E} \left[\left(V_{k_1, \dots, k_\ell}(G_N, s_1, \dots, s_\ell, D) V_{k'_1, \dots, k'_\ell}(G_N, s_1, \dots, s_\ell, D') \right)^{2p} \right] \\ \leq N^{2p} ((2p-1)!!)^2 \left(1 + \frac{(4kp)^{4kh}}{N^{1/3}} + \frac{(4kp)^{2kp}}{N^{(h+1)/3}} \right). \end{aligned} \quad (14)$$

Lemma 12. Since the proof is similar to the proof of [17, Lemma 3.8], we leave some details to the reader.

Expanding the left hand side of (14), we get that the expected value in (14) is

$$\begin{aligned} \sum_{n_1, \dots, n_{2p}=1}^{N - \lfloor \frac{N}{\log N} \rfloor} \sum_{m_1, \dots, m_{2p}=1}^{N - \lfloor \frac{N}{\log N} \rfloor} \mathbb{E} \left[\prod_{z=1}^{2p} \prod_{u=1}^{\ell} e_{n_z + d_1^u}(s_u) \dots e_{n_z + d_{k_u}^u}(s_u) \right. \\ \left. \cdot \prod_{v=1}^{\ell} e_{m_z + d_1'^v}(s_v) \dots e_{m_z + d_{k'_v}'^v}(s_v) \right]. \end{aligned} \quad (15)$$

Since $e_n(s)$ are mutually independent for $1 \leq n \leq N$ and $1 \leq s \leq S$ with $e_n(s) \in \{-1, 1\}$ and $\mathbb{E}[s_n(s)] = 0$, then (15) is the number of $4p$ -tuples $(n_1, \dots, n_{2p}, m_1, \dots, m_{2p})$ such that for each u the tuples $(n_z + d_1^u, \dots, n_z + d_{k_u}^u, m_z + d_1^u, \dots, m_z + d_{k_u}^u)_{z=1}^{2p}$ are even.

Then in the same way as in [17, Lemma 3.8] one may get that the number of such $4p$ -tuples is at most

$$((2p-1)!!N^p)^2 \left(1 + \frac{2p(k_i + k'_i)^{2h(k_i + k'_i)}}{N^{1/3}} + \frac{2p(k_i + k'_i)^{p(k_i + k'_i)}}{N^{(h+1)/3}} \right).$$

using Lemma 11. \square

Lemma 9. If X_1 and X_2 are random variables, then for all positive integers p and for $\theta_1, \theta_2 > 0$ Markov's inequality yields

$$\mathbb{P}[|X_1| \geq \theta_1 \cap |X_2| \geq \theta_2] \leq \frac{\mathbb{E}[(X_1 X_2)^{2p}]}{(\theta_1 \theta_2)^p}.$$

Let $p = \lfloor (k-1) \log N \rfloor$ and $h = \lfloor \alpha \log \log N \rfloor$ for some large $\alpha > 0$ to be fixed later.

By (14) and Markov's inequality we have

$$\begin{aligned} & \mathbb{P}[|V_{k_1, \dots, k_\ell}(G_N, s_1, \dots, s_\ell, D)| \geq \lambda \cap |V_{k'_1, \dots, k'_\ell}(G_N, s_1, \dots, s_\ell, D')| \geq \lambda] \\ & \leq \frac{\mathbb{E} \left[\left(V_{k_1, \dots, k_\ell}(G_N, s_1, \dots, s_\ell, D) V_{k'_1, \dots, k'_\ell}(G_N, s_1, \dots, s_\ell, D') \right)^{2p} \right]}{\lambda^{4p}} \\ & \leq \frac{((2p-1)!!)^2 N^{2p}}{(2N(k-1) \log N)^{2p}} \left(1 + \frac{(4kp)^{4kh}}{N^{1/3}} + \frac{(4kp)^{2kp}}{N^{(h+1)/3}} \right) \\ & = \frac{((2p-1)!!)^2}{(2(k-1) \log N)^{2p}} (1 + K_1(p, h) + K_2(p, h)), \end{aligned}$$

where

$$K_1(p, h) = \frac{(4kp)^{4kh}}{N^{1/3}}, \quad K_2(p, h) = \frac{(4kp)^{2kp}}{N^{(h+1)/3}}.$$

By Stirling's approximation (see e.g. [7]) it follows that

$$\sqrt{2\pi n} n^n e^{-n} \leq n! \leq \sqrt{3\pi n} n^n e^{-n}$$

so we have

$$\frac{((2p-1)!!)^2}{(2(k-1) \log N)^{2p}} \leq \frac{3e^2}{N^{2(k-1)}}.$$

Moreover

$$\begin{aligned} \log K_1(p, h) &= -\frac{1}{3} \log N + 4kh \log(4kp) \leq -\frac{1}{3} \log N + 6\alpha k \log k (\log \log N)^2 \\ &\leq -\frac{1}{4} \log N \end{aligned}$$

if N is large enough and

$$\begin{aligned}
\log K_2(p, h) &= -\frac{h+1}{3} \log N + 2kp \log(4kp) \\
&\leq -\frac{\alpha}{3} \log N \log \log N + 6k^2 \log k \log N \log \log N \\
&= \left(-\frac{\alpha}{3} + 6k^2 \log k\right) \log N \log \log N \\
&\leq -\log N \log \log N.
\end{aligned}$$

if we choose $\alpha = 10k^2 \log k$. Then the result follows. \square

Lemma 6. From Theorem 3 and Lemma 3 it follows that

$$\limsup_{N \rightarrow \infty} \frac{\mathbb{E}[\Phi_k(G_{N,S})]}{\sqrt{2N \log \left(\binom{S \cdot N}{k} - \binom{S \cdot (N-1)}{k} \right)}} \leq 1. \quad (16)$$

Now

$$\begin{aligned}
&\binom{S \cdot N}{k} - \binom{S \cdot (N-1)}{k} \\
&= \binom{S \cdot N}{k} \left(1 - \left(1 - \frac{k}{S \cdot N}\right) \cdots \left(1 - \frac{k}{S \cdot N - S + 1}\right)\right). \quad (17)
\end{aligned}$$

Since

$$\begin{aligned}
&\left(1 - \frac{k}{S \cdot N}\right) \cdots \left(1 - \frac{k}{S \cdot N - S + 1}\right) \\
&\leq \left(1 - \frac{k}{S \cdot N}\right)^S = 1 - \frac{k}{N} + O\left(\frac{k^2}{N^2}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\left(1 - \frac{k}{S \cdot N}\right) \cdots \left(1 - \frac{k}{S \cdot N - S + 1}\right) \\
&\geq \left(1 - \frac{k}{S(N-1)}\right)^S = 1 - \frac{k}{N-1} + O\left(\frac{k^2}{N^2}\right) = 1 - \frac{k}{N} + O\left(\frac{k}{N^2}\right),
\end{aligned}$$

we have that (17) is

$$\binom{S \cdot N}{k} - \binom{S \cdot (N-1)}{k} = \binom{S \cdot N}{k} \left(\frac{k}{N} + O\left(\frac{k}{N^2}\right)\right) \quad (18)$$

and by (6) its logarithm is

$$\log \binom{S \cdot N}{k} + \log \frac{k}{N} + \log \left(1 + O\left(\frac{1}{N}\right)\right) \sim (k-1) \log N$$

as S and k are fixed.

It follows from (16) that

$$\limsup_{N \rightarrow \infty} \frac{\mathbb{E} [\Phi_k (G_{N,S})]}{\sqrt{2N(k-1) \log N}} \leq 1$$

so it is enough to show that

$$\liminf_{N \rightarrow \infty} \frac{\mathbb{E} [\Phi_k (G_{N,S})]}{\sqrt{2N(k-1) \log N}} \geq 1.$$

Let $\delta > 0$ and put

$$N(\delta) = \left\{ N \geq k : \frac{\mathbb{E} [\Phi_k (G_{N,S})]}{\sqrt{2N(k-1) \log N}} < 1 - \delta \right\}.$$

We shall show, that $N(\delta)$ is a finite set for all $\delta > 0$ which proves the lemma according to (16).

Clearly,

$$\Phi_k (G_{N,S}) \geq \max_{\ell, k_1, \dots, k_\ell} \max_{s_1, \dots, s_\ell} \max_D |V_{k_1, \dots, k_\ell} (G_{N,S}, s_1, \dots, s_\ell, D)|,$$

so

$$\begin{aligned} \mathbb{P} [\Phi_k (G_{N,S}) \geq \lambda] &\geq \sum \mathbb{P} [|V_{k_1, \dots, k_\ell} (G_{N,S}, s_1, \dots, s_\ell, D)| \geq \lambda] \\ &\quad - \frac{1}{2} \sum \mathbb{P} [|V_{k_1, \dots, k_\ell} (G_{N,S}, s_1, \dots, s_\ell, D)| \geq \lambda \\ &\quad \cap |V_{k'_1, \dots, k'_{\ell'}} (G_{N,S}, s'_1, \dots, s'_{\ell'}, D')| \geq \lambda], \end{aligned} \quad (19)$$

where the first sum is taken over all positive integers ℓ and k_1, \dots, k_ℓ with $k_1 + \dots + k_\ell = k$, all distinct $1 \leq s_1, \dots, s_\ell \leq S$ and all D having the form (10), while the second sum is taken over all positive integers ℓ, ℓ' and $k_1, \dots, k_\ell, k'_1, \dots, k'_{\ell'}$ with $k_1 + \dots + k_\ell = k'_1 + \dots + k'_{\ell'} = k$, all distinct $1 \leq s_1, \dots, s_\ell \leq S$ and distinct $1 \leq s'_1, \dots, s'_{\ell'} \leq S$ and all D, D' having the form (10) with the additional restriction that $D \neq D'$ if $\{s_1, \dots, s_\ell\} = \{s'_1, \dots, s'_{\ell'}\}$.

First we give a lower bound to the first term of (19) by Lemma 7. Namely,

$$\begin{aligned} &\sum_{\ell=1}^k \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_D \mathbb{P} [|V_{k_1, \dots, k_\ell} (G_{N,S}, s_1, \dots, s_\ell, D)| \geq \lambda] \\ &\geq \frac{1}{2e^{k-1} N^{k-1} \sqrt{\pi(k-1) \log N}} \sum_{\ell=1}^k \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_D 1 \\ &= \left(\binom{S \cdot (\lfloor N/\log N \rfloor + 1)}{k} - \binom{S \cdot \lfloor N/\log N \rfloor}{k} \right) \frac{1}{2e^{k-1} N^{k-1} \sqrt{\pi(k-1) \log N}} \\ &\geq \binom{S \cdot (\lfloor N/\log N \rfloor + 1)}{k} \cdot \frac{k}{N} \frac{1}{4e^{k-1} N^{k-1} \sqrt{\pi(k-1) \log N}} \\ &\geq \frac{S^k}{4(ek)^{k-1} (\log N)^k \sqrt{\pi(k-1) \log N}} \end{aligned} \quad (20)$$

using (6) and (18) with N replaced by $\lfloor N/\log N \rfloor + 1$.

We give a lower bound to the second term of (19) by Lemmas 7 and 9. If $\{s_1, \dots, s_\ell\} \neq \{s'_1, \dots, s'_{\ell'}\}$, then $V_{k_1, \dots, k_\ell}(G_{N,S}, s_1, \dots, s_\ell, D)$ and $V_{k'_1, \dots, k'_{\ell'}}(G_{N,S}, s'_1, \dots, s'_{\ell'}, D')$ are independent by Lemma 5, thus by Lemma 7 we have in the same way that

$$\begin{aligned}
& \sum_{\ell, \ell'=1}^k \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_{1 \leq s'_1, \dots, s'_{\ell'} \leq S} \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_{\substack{k'_1, \dots, k'_{\ell'} \geq 1 \\ k'_1 + \dots + k'_{\ell'} = k}} \sum_{D, D'} \\
& \quad \mathbb{P}[|V_{k_1, \dots, k_\ell}(G_{N,S}, s_1, \dots, s_\ell, D)| \geq \lambda \cap |V_{k'_1, \dots, k'_{\ell'}}(G_{N,S}, s'_1, \dots, s'_{\ell'}, D')| \geq \lambda] \\
& \leq \sum_{\ell, \ell'=1}^k \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_{1 \leq s'_1, \dots, s'_{\ell'} \leq S} \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_{\substack{k'_1, \dots, k'_{\ell'} \geq 1 \\ k'_1 + \dots + k'_{\ell'} = k}} \sum_{D, D'} \\
& \quad \cdot \frac{2}{e^{2(k-1)} N^{2(k-1)} \pi(k-1) \log N} \\
& = \left(\binom{S \cdot (\lfloor N/\log N \rfloor + 1)}{k} - \binom{S \cdot \lfloor N/\log N \rfloor}{k} \right)^2 \\
& \quad \cdot \frac{2}{e^{2(k-1)} N^{2(k-1)} \pi(k-1) \log N} \\
& \leq \left(\binom{S \cdot (\lfloor N/\log N \rfloor + 1)}{k} \cdot \frac{k}{N} \right)^2 \frac{4}{e^{2(k-1)} N^{2(k-1)} \pi(k-1) \log N} \\
& \leq \frac{4S^{2k}}{k^{2(k-1)} (\log N)^{2k} \pi(k-1) \log N} \tag{21}
\end{aligned}$$

by (6).

For $\{s_1, \dots, s_\ell\} = \{s'_1, \dots, s'_{\ell'}\}$ we have by Lemma 9, that

$$\begin{aligned}
& \sum_{\ell=1}^k \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_{\substack{k'_1, \dots, k'_{\ell'} \geq 1 \\ k'_1 + \dots + k'_{\ell'} = k}} \sum_{D, D'} \\
& \quad \mathbb{P}[|V_{k_1, \dots, k_\ell}(G_{N,S}, s_1, \dots, s_\ell, D)| \geq \lambda \cap |V_{k'_1, \dots, k'_{\ell'}}(G_{N,S}, s'_1, \dots, s'_{\ell'}, D')| \geq \lambda] \\
& \leq \sum_{\ell=1}^k \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_{\substack{k'_1, \dots, k'_{\ell'} \geq 1 \\ k'_1 + \dots + k'_{\ell'} = k}} \sum_{D, D'} \frac{23}{N^{2(k-1)}} \tag{22}
\end{aligned}$$

Since one of the $d_1^{s_i}$ in D' takes the value 0, for fixed ℓ and s_1, \dots, s_ℓ , the number of possible ℓ -tuples $(k'_1, \dots, k'_{\ell'})$ and D' is at most

$$\ell \cdot \binom{\ell \cdot (\lfloor N/\log N \rfloor + 1)}{k-1} \leq k \binom{k \cdot (\lfloor N/\log N \rfloor + 1)}{k-1} \leq \frac{k^k 2^{k-1} N^{k-1}}{(k-1)^{k-1} (\log N)^{k-1}}$$

by (6). Thus (22) is at most

$$\begin{aligned}
& \frac{k^k 2^{k-1} N^{k-1}}{(k-1)^{k-1} (\log N)^{k-1}} \sum_{\ell=1}^k \sum_{1 \leq s_1, \dots, s_\ell \leq S} \sum_{\substack{k_1, \dots, k_\ell \geq 1 \\ k_1 + \dots + k_\ell = k}} \sum_D \frac{23}{N^{2(k-1)}} \\
& \leq \frac{k^k 2^{k-1} N^{k-1}}{(k-1)^{k-1} (\log N)^{k-1}} \frac{23}{N^{2(k-1)}} \\
& \quad \cdot \left(\binom{S \cdot (\lfloor N/\log N \rfloor + 1)}{k} - \binom{S \cdot \lfloor N/\log N \rfloor}{k} \right)^2 \\
& \leq \frac{23 k e^k 2^{2k-1} S^k}{(k-1)^{k-1} (\log N)^{2k-1}} \tag{23}
\end{aligned}$$

by (6).

As $N \rightarrow \infty$, (20) dominates (21) and (23) thus we have from (19) that

$$\mathbb{P}(\Phi_k(G_{N,S}) \geq \lambda) \geq \frac{S^k}{5(ek)^{k-1} (\log N)^k \sqrt{\pi(k-1) \log N}} \tag{24}$$

By the definition of $N(\delta)$, we have $\lambda > \mathbb{E}(\Phi_k(G_{N,S}))$ for $N \in N(\delta)$, thus by Lemma 3 we have for $\theta = \lambda - \mathbb{E}(\Phi_k(G_{N,S}))$ that

$$\mathbb{P}[\Phi_k(G_{N,S}) \geq \lambda] \leq 2 \exp \left\{ -\frac{(\lambda - \mathbb{E}(\Phi_k(G_{N,S})))^2}{2k^2 N} \right\}$$

for $N \in N(\delta)$. Comparing it with (24) we get

$$\frac{S^k}{5(ek)^{k-1} (\log N)^k \sqrt{\pi((k-1) \log N + k \log S)}} \leq 2 \exp \left\{ -\frac{(\lambda - \mathbb{E}(\Phi_k(G_{N,S})))^2}{2k^2 N} \right\}$$

i.e.

$$\begin{aligned}
& \frac{\mathbb{E}(\Phi_k(G_{N,S}))}{\sqrt{2N(k-1) \log N}} \\
& \geq 1 - \sqrt{k^2 \frac{\log(10(ek)^{k-1} \pi^{1/2}) + k \log \log N + \frac{1}{2} \log(k-1) \log N - k \log S}{(k-1) \log N}}
\end{aligned}$$

Since the right hand side goes to 1 as $N \rightarrow \infty$, we see that the size of $N(\delta)$ is finite. □

Let

$$S^\pm(n) = \sum_{1 \leq i \leq n} X_i,$$

where X_i ($1 \leq i \leq n$) are independent random variables with mean 0, that is,

$$\mathbb{P}(X_i = -1) = \mathbb{P}(X_i = +1) = 1/2.$$

The following lemma states a well-known estimate for large deviation of $S^\pm(n)$ (see, e.g. [4, Appendix 2]):

Lemma 13. Let X_i ($1 \leq i \leq n$) be independent ± 1 random variables with mean 0. Let $S^\pm(n) = \sum_{1 \leq i \leq n} X_i$. For any real number $a > 0$, we have

$$\mathbb{P}(S^\pm(n) > a) < e^{-a^2/2n}.$$

Lemma 14. Let G_S be drawn from Ω equipped with the probability measure defined by (2) with $G_S(s) = (e_1(s), e_2(s), \dots)$ for $1 \leq s \leq S$. Let $G_N(s) = (e_1(s), \dots, e_N(s))$ for $1 \leq s \leq S$. Let N_1, N_2, \dots be a strictly increasing sequence of integers. Then, almost surely

$$\Phi_k(G_{N_{r+1}, S}) - \Phi_k(G_{N_r, S}) \leq \sqrt{6(N_{r+1} - N_r)(k-1) \log N_{r+1}}$$

for all sufficiently large r .

Proof. Write

$$\lambda = \sqrt{6(N_{r+1} - N_r)(k-1) \log N_{r+1}}.$$

If

$$\Phi_k(G_{N_{r+1}, S}) - \Phi_k(G_{N_r, S}) > \lambda, \quad (25)$$

then there is a tuple (d_1, \dots, d_k) with

$$0 \leq d_1 \leq \dots \leq d_k < N_{r+1}, \quad (26)$$

an integer m with

$$N_r - d_k + 1 \leq m \leq N_{r+1} - d_k, \quad (27)$$

and $1 \leq s_1, \dots, s_k \leq S$ such that

$$\left| \sum_{n=\max\{1, N_r - d_k + 1\}}^m e_{n+d_1}(s_1) \dots e_{n+d_k}(s_k) \right| > \lambda. \quad (28)$$

By Lemmas 5 and 13 we have that the probability, that (28) holds is at most

$$\exp \left\{ -\frac{\lambda^2}{2(N_{r+1} - N_r)} \right\} \leq N_{r+1}^{-3k+3}.$$

Summing up all possible tuples (d_1, \dots, d_k) , all possible integers m and all possible $1 \leq s_1, \dots, s_k \leq S$, the probability that (28) happens for some (d_1, \dots, d_k) satisfying (26), some m satisfying (27) and some $1 \leq s_1, \dots, s_k \leq S$ is at most

$$(N_{r+1} - N_r)(N_{r+1}S)^k N_{r+1}^{-3k+3} \leq N_{r+1}^{-2k+4} S^k.$$

This is also an upper bound for the probability of (25), and so

$$\mathbb{P}[\Phi_k(G_{N_{r+1}}) - \Phi_k(G_{N_r}) > \lambda] \leq N_{r+1}^{-2k+4} S^k.$$

Summing it over all r we get

$$\sum_{r=1}^{\infty} \frac{1}{2k-1} N_{r+1}^{-2k+4} S^k < \infty$$

and the result follows from the Borel-Cantelli Lemma. \square

Theorem 4. Write

$$\lambda(N) = \sqrt{2kN \log N}.$$

Let $N_r = \lceil e^{r^{1/2}} \rceil$ ($r = 1, 2, \dots$) be a sequence of integers. We remark, that

$$\lim_{k \rightarrow \infty} \frac{N_{r+1}}{N_r} = \lim_{k \rightarrow \infty} e^{(r+1)^{1/2} - r^{1/2}} = 1.$$

First we prove, that

$$\max_{N_{r-1} \leq N \leq N_r} \left| \frac{\Phi_k(G_{N,S})}{\lambda(N_r)} - 1 \right| \rightarrow 0 \quad \text{almost surely.} \quad (29)$$

For any $\varepsilon > 0$ we have

$$\begin{aligned} & \mathbb{P} \left[\left| \frac{\Phi_k(G_{N_r,S})}{\lambda(N_r)} - 1 \right| > \varepsilon \right] \\ & \leq \mathbb{P} \left[\left| \frac{\mathbb{E}[\Phi_k(G_{N_r,S})]}{\lambda(N_r)} - 1 \right| > \frac{\varepsilon}{2} \right] + \mathbb{P} \left[\left| \frac{\Phi_k(G_{N_r,S})}{\lambda(N_r)} - \frac{\mathbb{E}[\Phi_k(G_{N_r,S})]}{\lambda(N_r)} \right| > \frac{\varepsilon}{2} \right]. \end{aligned}$$

The first term equals zero for sufficiently large r by Lemma 6. By Lemma 3 the second term is at most

$$2 \exp \left\{ -\frac{\varepsilon^2 \lambda(N_r)^2}{8k^2 N_r} \right\} \leq \exp \left\{ -\frac{\varepsilon^2 \log N_r}{4k} \right\} \leq \exp \left\{ -\frac{\varepsilon^2 \log N_r}{4k} \right\}.$$

Applying a crude estimate we get that for sufficiently large r

$$\exp \left\{ -\frac{\varepsilon^2 \log N_r}{4k} \right\} \leq \frac{1}{4k \log N_r^3} \leq r^{-3/2}.$$

Thus for a sufficiently large r_0 we have

$$\sum_{r=r_0}^{\infty} \mathbb{P} \left[\left| \frac{\Phi_k(G_{N_r,S})}{\lambda(N_r)} - 1 \right| > \varepsilon \right] \leq \sum_{r=r_0}^{\infty} r^{-3/2} < \infty$$

and (29) follows from the Borel-Cantelli Lemma.

Next we show

$$\max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N,S})}{\lambda(N)} - 1 \right| \rightarrow 0 \quad \text{almost surely.}$$

By the triangle inequality we have

$$\begin{aligned} & \max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N,S})}{\lambda(N)} - 1 \right| \\ & \leq \max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N_{r+1},S})}{\lambda(N_{r+1})} - 1 \right| + \max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N_{r+1},S})}{\lambda(N_{r+1})} - \frac{\Phi_k(G_{N,S})}{\lambda(N_{r+1})} \right| \\ & + \max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N,S})}{\lambda(N)} - \frac{\Phi_k(G_{N,S})}{\lambda(N_{r+1})} \right|. \end{aligned} \quad (30)$$

The first term goes to zero almost surely by (29). Since $\Phi_k(G_{N,S})$ is non-decreasing in N , we have by Lemma 14 that for sufficiently large r

$$\max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N_{r+1},S})}{\lambda(N_{r+1})} - \frac{\Phi_k(G_{N,S})}{\lambda(N_{r+1})} \right| \leq \sqrt{3 \frac{N_{r+1} - N_r}{N_{r+1}}} \rightarrow 0. \quad (31)$$

On the other hand

$$\begin{aligned} \max_{N_r \leq N \leq N_{r+1}} \left| \frac{\Phi_k(G_{N,S})}{\lambda(N)} - \frac{\Phi_k(G_{N,S})}{\lambda(N_{r+1})} \right| \\ \leq \frac{\Phi_k(G_{N_{r+1}})}{\lambda(N_{r+1})} \max_{N_r \leq N \leq N_{r+1}} \left| \frac{\lambda(N_{r+1})}{\lambda(N)} - 1 \right| \rightarrow 0, \quad \text{almost surely} \end{aligned} \quad (32)$$

by (29).

Finally, the result follows from (29), (30), (31) and (32). \square

4 Minimal values of $\Phi_k(G)$

First we prove Theorem 5.

Theorem 5. For the lower bound we can assume, that $4k+1 < S$ otherwise the bound is trivial. Consider the maximal integer L such that

$$2k \cdot 2^L + 1 \leq S.$$

By the pigeon hole principle, there are different numbers $1 \leq s_1 < s_2 < \dots < s_{2k+1} \leq S$ such that their first L elements are coincide:

$$e_n(s_1) = e_n(s_2) = \dots = e_n(s_{2k+1}), \quad n = 1, 2, \dots, L.$$

Thus

$$\Phi_{2k+1}(G_{N,S}) \geq \left| \sum_{n=1}^L e_n(s_1) e_n(s_2) \dots e_n(s_{2k+1}) \right| = L = \left\lfloor \log_2 \frac{S-1}{2k} \right\rfloor.$$

For the upper bound, given S we construct a map $G_{N,S}$ with small cross-correlation measure $\Phi_{2k+1}(G_{N,S})$.

Let $K = \lceil \log_2 S \rceil$ and define the sequences by

$$e_n(i) = \begin{cases} (-1)^{i_n}, & 1 \leq n < K, \\ (-1)^{i_K+n}, & K \leq n \leq N, \end{cases} \quad i = 1, 2, \dots, S$$

where $i_1, i_2, \dots, i_K \in \{0, 1\}$ are the binary digits of $i-1$

$$i-1 = \sum_{n=1}^K i_n 2^{n-1}.$$

Then for $M \leq N$, all $0 \leq d_1 \leq \dots \leq d_{2k+1} \leq N - M$ and all $1 \leq s_1, s_2, \dots, s_{2k+1} \leq S$ with $d_i \neq d_j$ if $s_i = s_j$ we have

$$\begin{aligned} & \left| \sum_{n=1}^M e_{n+d_1}(s_1) e_{n+d_2}(s_2) \dots e_{n+d_{2k+1}}(s_{2k+1}) \right| \\ & \leq \left| \sum_{n=1}^{K-1} e_{n+d_1}(s_1) e_{n+d_2}(s_2) \dots e_{n+d_{2k+1}}(s_{2k+1}) \right| \\ & \quad + \left| \sum_{n=K}^M e_{n+d_1}(s_1) e_{n+d_2}(s_2) \dots e_{n+d_{2k+1}}(s_{2k+1}) \right| \\ & \leq K - 1 + \left| \sum_{n=K}^M (-1)^n \right| \leq K \end{aligned}$$

which proves the upper bound. \square

The proof of Theorem 6 is similar to the proof of (1), it is based on the following lemmas (Lemmas 5 and 6 in [2]).

Lemma 15. For a symmetric matrix $\mathbf{A} = (A_{i,j})_{1 \leq i,j \leq n}$ with $A_{i,i} = 1$ for all i and $|A_{i,j}| \leq \varepsilon$ for all $i \neq j$ we have

$$\text{rk}(\mathbf{A}) \geq \frac{n}{1 + \varepsilon^2(n-1)}.$$

Lemma 16. Let $\mathbf{A} = (A_{i,j})_{1 \leq i,j \leq n}$ be a real matrix with $A_{i,i} = 1$ for all i and $|A_{i,j}| \leq \varepsilon$ for all $i \neq j$, where $\sqrt{1/n} \leq \varepsilon \leq 1/2$. Then

$$\text{rk}(\mathbf{A}) \geq \frac{1}{100\varepsilon^2 \log(1/\varepsilon)} \log n. \quad (33)$$

If \mathbf{A} is symmetric, then (33) holds with the constant $1/100$ replaced by $1/50$.

Theorem 6. First consider the case when S is large $2kN \leq S$. Let $t = \lfloor S/k \rfloor$ and let $L_1, L_2, \dots, L_t \subset \{1, 2, \dots, S\}$ be distinct subsets with k elements. For each L_i ($i = 1, 2, \dots, t$) we assign the vector $\mathbf{v}_i \in \mathbb{R}^N$ with

$$\mathbf{v}_{i,j} = \prod_{s \in L_i} e_j(s).$$

Define the matrix $\mathbf{A} = (A_{i,j})_{1 \leq i,j \leq t}$ by

$$A_{i,j} = \frac{1}{N} \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \frac{1}{N} \sum_{n=1}^N \prod_{s \in L_i \cup L_j} e_n(s),$$

where $\langle \cdot, \cdot \rangle$ is the standard inner-product. Then we have

$$A_{i,i} = 1, \quad i = 1, 2, \dots, t,$$

and

$$\Phi_{2k}(G_{N,S}) \geq N \max\{|A_{i,j}| : 1 \leq i, j \leq t, i \neq j\}. \quad (34)$$

Let $\mathbf{B} = (\mathbf{v}_i^T)_{1 \leq i \leq t}$ be the $t \times N$ matrix with rows \mathbf{v}_i^T ($1 \leq i \leq t$). Clearly $\mathbf{A} = N^{-1}\mathbf{B}\mathbf{B}^T$, thus $\text{rk}(\mathbf{A}) \leq N$. On the other hand, by Lemma 15 we have

$$N \geq \text{rk}(\mathbf{A}) \geq \frac{t}{1 + \varepsilon^2(t-1)}, \quad (35)$$

where $\varepsilon = \max\{|A_{i,j}| : 1 \leq i, j \leq t, i \neq j\}$. (35) gives

$$\varepsilon^2 \geq \frac{1}{N} - \frac{1}{t} \geq \frac{1}{t}. \quad (36)$$

Now we are ready to complete the proof of the result. If $\varepsilon > 1/2$, then the theorem trivially holds. Otherwise, by (36) we have $\sqrt{1/t} \leq \varepsilon \leq 1/2$, thus by Lemma 16 we have

$$N \geq \text{rk}(\mathbf{A}) \geq \frac{1}{50\varepsilon^2 \log(1/\varepsilon)} \log \left\lfloor \frac{S}{k} \right\rfloor,$$

whence

$$\varepsilon^2 \log(1/\varepsilon) \geq \frac{1}{50N} \log \left\lfloor \frac{S}{k} \right\rfloor. \quad (37)$$

Using that $1/\varepsilon \geq \log 1/\varepsilon$, we have from (37), that

$$\varepsilon \geq \varepsilon^2 \log(1/\varepsilon) \geq \frac{1}{50N} \log \left\lfloor \frac{S}{k} \right\rfloor. \quad (38)$$

Thus from (37) and (38) we get

$$\varepsilon^2 \log \frac{50N}{\log \lfloor S/k \rfloor} \geq \frac{1}{50N} \log \left\lfloor \frac{S}{k} \right\rfloor,$$

and hence

$$\varepsilon \geq \sqrt{\frac{\log \lfloor S/k \rfloor}{50N}} \bigg/ \log \frac{50N}{\log \lfloor S/k \rfloor}. \quad (39)$$

Then (3) follows from (34) and (39).

Next, consider the case when S is small $2kN > S$. Put $\ell = \lceil k/S \rceil$ and write $M = \lfloor N/(2\ell+1) \rfloor$ and $N' = N - M + 1$. Let L_1, \dots, L_t pairwise disjoint k -element subsets of $\{1, \dots, S\} \times \{1, \dots, N'\}$ with

$$t \geq S \cdot \left\lfloor \frac{N'}{\ell} \right\rfloor.$$

Then

$$t \geq S \cdot \left\lfloor \frac{N - \lfloor N/(2\ell+1) \rfloor + 1}{\ell} \right\rfloor \geq 2M.$$

In the same way as before, we get from Lemma 15 that

$$\Phi_{2k} \geq M \sqrt{\frac{1}{M} - \frac{1}{t}} \geq \sqrt{M - \frac{M^2}{t}} \geq \sqrt{M - \frac{M}{2}} \geq \sqrt{\frac{N}{2\lceil k/S \rceil + 1}}$$

and (4) follows. \square

Acknowledgment

The author would like to thank András Sárközy for helpful discussions. The author also thank the anonymous referees for their valuable comments.

The author is partially supported by the Austrian Science Fund FWF Project F5511-N26 which is part of the Special Research Program "Quasi-Monte Carlo Methods: Theory and Applications" and by Hungarian National Foundation for Scientific Research, Grant No. K100291.

References

- [1] R. Ahlswede, J. Cassaigne, A Sárközy, On the correlation of binary sequences, *Discrete Appl. Math.* **156** (2008), no. 9, 1478–1487.
- [2] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira, V. Rödl, Measures of pseudorandomness for finite sequences: minimal values, *Combin. Probab. Comput.* **15** (2006), no. 1-2, 1–29.
- [3] N. Alon, Y. Kohayakawa, C. Mauduit, C. G. Moreira, V. Rödl, Measures of pseudorandomness for finite sequences: typical values, *Proc. Lond. Math. Soc.* **95** (2007), no. 3, 778–812.
- [4] N. Alon, J. H. Spencer, *The probabilistic method*, second ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience [John Wiley & Sons], New York, 2000, With an appendix on the life and work of Paul Erdős.
- [5] J. Cassaigne, C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences VII: The measures of pseudorandomness, *Acta Arith.* **103** (2002), no. 2, 97–118.
- [6] B. Bollobás, *Random graphs*, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1985.
- [7] J. W. Feller, *An introduction to probability theory and its applications*. Vol. I, Third edition, John Wiley & Sons Inc., New York, 1968.
- [8] K. Gyarmati, On the correlation of binary sequences, *Studia Sci. Math. Hungar.* **42** (2005), 59–75
- [9] K. Gyarmati, Measures of pseudorandomness, P. Charpin, A. Pott, A. Winterhof (eds.), *Radon Series in Computational and Applied Mathematics*, de Gruyter 2013, 43–64.
- [10] K. Gyarmati, C. Mauduit, A. Sárközy, The cross-correlation measure for families of binary sequences, *Applied algebra and number theory*, 126–143, Cambridge Univ. Press, Cambridge, 2014.
- [11] C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences I: Measures of pseudorandomness, the Legendre symbol, *Acta Arith.* **82** (1997) 365–377.
- [12] C. Mauduit, A. Sárközy, On the measures of pseudorandomness of binary sequences, *Discrete Math.* **271** (2003), no. 1-3, 195–207.

- [13] C. McDiarmid, On the method of bounded differences, Surveys in combinatorics, 1989 (Norwich, 1989), London Math. Soc. Lecture Note Ser., vol. 141, Cambridge Univ. Press, Cambridge, 1989, pp. 148–188.
- [14] L. Mérai, On the typical values of the cross-correlation measure, Monatsh. Math. **180** (2016) no. 1 83–99.
- [15] A. Sárközy, On pseudorandomness of families of binary sequences, Discrete Applied Mathematics, (2015) doi:10.1016/j.dam.2015.07.031
- [16] K.-U. Schmidt, The peak sidelobe level of random binary sequences, Bull. London Math. Soc. **46** (2014), no. 3, 643–652.
- [17] K.-U. Schmidt, The correlation measures of finite sequences: limiting distributions and minimum values, Trans. Amer. Math. Soc., (2016) doi: 10.1090/tran6650